MIS510 Portfolio Project Option 1

Krista ONeill

10/2/2020

Data Exploration

```
#Calling required libraries for the analyses performed.
library(gains)
library(e1071)
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

#Reading GermanCredit.csv into R and then creating a dataframe with summary s
tatistics.
#There are no missing values for any of the observations.

gcredit <- read.csv("GermanCredit.csv", header=TRUE)

#Creating a dataframe with summary values for each numerical column.
gcredit.summary <- data.frame(mean=sapply(gcredit[,],mean,na.rm=TRUE),
sd=sapply(gcredit[,],sd,na.rm=TRUE),
min=sapply(gcredit[,],min,na.rm=TRUE),
max=sapply(gcredit[,],max,na.rm=TRUE),
median=sapply(gcredit[,],median,na.rm=TRUE),
length=sapply(gcredit[,],length),
miss.val=sapply(gcredit[,],function(x) sum(length(which(is.na(x))))))


gcredit.summary

##                      mean           sd min  max median length miss.val
## OBS.              500.500  288.8194361   1 1000  500.5   1000        0
## CHK_ACCT            1.577    1.2576377   0    3    1.0   1000        0
```

```
## DURATION            20.903   12.0588145   4    72   18.0   1000         0
## HISTORY              2.545    1.0831196   0     4    2.0   1000         0
## NEW_CAR              0.234    0.4235840   0     1    0.0   1000         0
## USED_CAR             0.103    0.3041110   0     1    0.0   1000         0
## FURNITURE            0.181    0.3852108   0     1    0.0   1000         0
## RADIO.TV             0.280    0.4492236   0     1    0.0   1000         0
## EDUCATION            0.050    0.2180540   0     1    0.0   1000         0
## RETRAINING           0.097    0.2961059   0     1    0.0   1000         0
## AMOUNT            3271.258 2822.7368760 250 18424 2319.5   1000         0
## SAV_ACCT             1.105    1.5800226   0     4    0.0   1000         0
## EMPLOYMENT           2.384    1.2083063   0     4    2.0   1000         0
## INSTALL_RATE         2.973    1.1187147   1     4    3.0   1000         0
## MALE_DIV             0.050    0.2180540   0     1    0.0   1000         0
## MALE_SINGLE          0.548    0.4979397   0     1    1.0   1000         0
## MALE_MAR_or_WID      0.092    0.2891706   0     1    0.0   1000         0
## CO.APPLICANT         0.041    0.1983894   0     1    0.0   1000         0
## GUARANTOR            0.052    0.2221381   0     1    0.0   1000         0
## PRESENT_RESIDENT     2.845    1.1037179   1     4    3.0   1000         0
## REAL_ESTATE          0.282    0.4501985   0     1    0.0   1000         0
## PROP_UNKN_NONE       0.154    0.3611294   0     1    0.0   1000         0
## AGE                 35.546   11.3754686  19    75   33.0   1000         0
## OTHER_INSTALL        0.186    0.3893014   0     1    0.0   1000         0
## RENT                 0.179    0.3835441   0     1    0.0   1000         0
## OWN_RES              0.713    0.4525879   0     1    1.0   1000         0
## NUM_CREDITS          1.407    0.5776545   1     4    1.0   1000         0
## JOB                  1.904    0.6536140   0     3    2.0   1000         0
## NUM_DEPENDENTS       1.155    0.3620858   1     2    1.0   1000         0
## TELEPHONE            0.404    0.4909430   0     1    0.0   1000         0
## FOREIGN              0.037    0.1888562   0     1    0.0   1000         0
## RESPONSE             0.700    0.4584869   0     1    1.0   1000         0
```

```r
#Discovering the range of both AGE and AMOUNT.
range(gcredit$AGE)
```

```
## [1] 19 75
```

```r
range(gcredit$AMOUNT)
```

```
## [1]   250 18424
```

Partitioning Data

```r
#Creating training & validation sets for the data. 60% of data is used for tr
aining and 40% for validation.

train.index <- sample(c(1:dim(gcredit)[1]), dim(gcredit)[1]*0.6)

gcredit.train <- gcredit[train.index,]
gcredit.valid <- gcredit[-train.index,]
```

Partitioning Data for Logistic Regression

```
#Choosing columns for logistic regression.  Column names & numbers are procur
ed via data.frame(colnames).

data.frame(colnames(gcredit))

##    colnames.gcredit.
## 1               OBS.
## 2           CHK_ACCT
## 3           DURATION
## 4            HISTORY
## 5            NEW_CAR
## 6           USED_CAR
## 7          FURNITURE
## 8           RADIO.TV
## 9          EDUCATION
## 10        RETRAINING
## 11            AMOUNT
## 12          SAV_ACCT
## 13        EMPLOYMENT
## 14      INSTALL_RATE
## 15          MALE_DIV
## 16       MALE_SINGLE
## 17   MALE_MAR_or_WID
## 18      CO.APPLICANT
## 19         GUARANTOR
## 20  PRESENT_RESIDENT
## 21       REAL_ESTATE
## 22     PROP_UNKN_NONE
## 23               AGE
## 24     OTHER_INSTALL
## 25              RENT
## 26           OWN_RES
## 27       NUM_CREDITS
## 28               JOB
## 29    NUM_DEPENDENTS
## 30         TELEPHONE
## 31           FOREIGN
## 32          RESPONSE

gcredit.train <- gcredit.train[,c(2,3,4,9,11,13,14,23,27,28,29,32)]

gcredit.valid <- gcredit.valid[,c(2,3,4,9,11,13,14,23,27,28,29,32)]

#In order to prepare for the logistic regression, which will be measuring non
-categorical variables only, we will take only DURATION, AGE, AMOUNT, NUM_DEP
ENDENT from the original training & validation sets.  RESPONSE is also includ
ed.

gcredit.train.log <- gcredit.train[,c(2,3,4,5,8,11,12)]
```

```r
gcredit.valid.log <- gcredit.valid[,c(2,3,4,5,8,11,12)]

#Setting both HISTORY and EDUCATION as factors.

gcredit.train.log$HISTORY<- factor(gcredit.train.log$HISTORY)
gcredit.train.log$EDUCATION <- factor(gcredit.train.log$EDUCATION)

gcredit.valid.log$HISTORY <- factor(gcredit.valid.log$HISTORY)
gcredit.valid.log$EDUCATION <- factor(gcredit.valid.log$EDUCATION)
```

Performing Logistic Regression

```r
#Logistic regression is used to determine the predictors' effect on the depen
dent variable, RESPONSE.

gcredit.lreg <- glm(RESPONSE~., data=gcredit.train.log, family="binomial")

options=(scipen=999)

summary(gcredit.lreg)

##
## Call:
## glm(formula = RESPONSE ~ ., family = "binomial", data = gcredit.train.log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2551  -1.1063   0.6333   0.8248   1.7640
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.756e-01  6.436e-01  -0.739 0.459904
## DURATION        -3.876e-02  9.992e-03  -3.879 0.000105 ***
## HISTORY1         9.574e-02  5.832e-01   0.164 0.869605
## HISTORY2         1.103e+00  4.611e-01   2.391 0.016801 *
## HISTORY3         1.278e+00  5.346e-01   2.390 0.016854 *
## HISTORY4         1.888e+00  4.897e-01   3.856 0.000115 ***
## EDUCATION1      -1.049e+00  4.011e-01  -2.615 0.008911 **
## AMOUNT           3.717e-05  4.426e-05   0.840 0.401041
## AGE              1.905e-02  9.489e-03   2.008 0.044683 *
## NUM_DEPENDENTS   2.173e-01  2.696e-01   0.806 0.420250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 734.72  on 599  degrees of freedom
## Residual deviance: 667.15  on 590  degrees of freedom
## AIC: 687.15
```

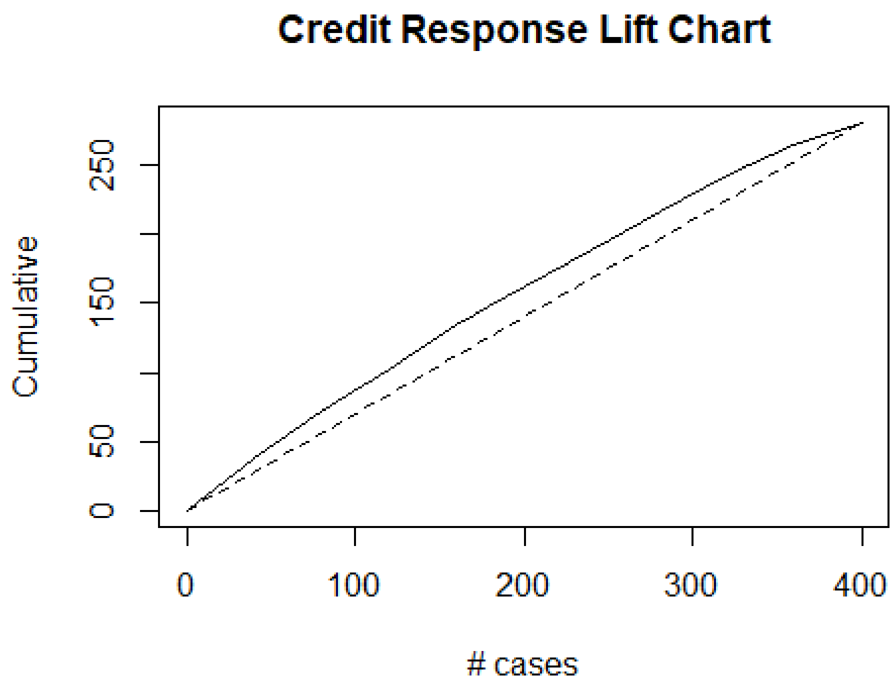
```
##
## Number of Fisher Scoring iterations: 4
```

The variables DURATION and HISTORY4 are marked with 3 stars, indicating that their p-values are very low and they are very significant to whether or not RESPONSE = 1 and credit is good.

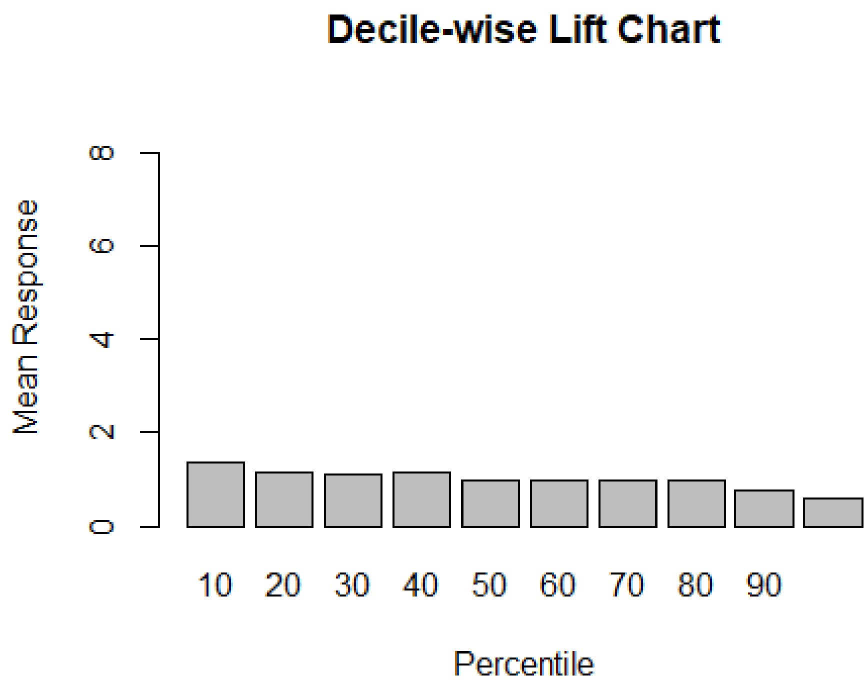
HISTORY2 and AGE have 1 star, meaning their p-values are still, if less, significant, and they do have a statistically significant impact on whether or not RESPONSE = 1.

Lift Chart

```
#In order to evaluate the effectiveness of the predictive model, a linear lif
t chart is created.

#Additionally, a decile-wise lift chart is created.

gcredit.lreg.pred <- predict(gcredit.lreg, gcredit.valid.log, type="response"
)

#Finding the gains from the validation set.

gcredit.gains <- gains(gcredit.valid.log$RESPONSE, gcredit.lreg.pred)

plot(c(0,gcredit.gains$cume.pct.of.total*sum(gcredit.valid.log$RESPONSE))~c(0
,gcredit.gains$cume.obs), xlab="# cases", ylab="Cumulative", main="Credit Res
ponse Lift Chart", type="l")

lines(c(0,sum(gcredit.valid.log$RESPONSE))~c(0, dim(gcredit.valid.log)[1]), l
ty=2)
```




```
heights <- gcredit.gains$mean.resp/mean(gcredit.valid.log$RESPONSE)

midpoints <- barplot(heights, names.arg = gcredit.gains$depth, ylim=c(0,9), x
lab="Percentile", ylab="Mean Response", main="Decile-wise Lift Chart")
```

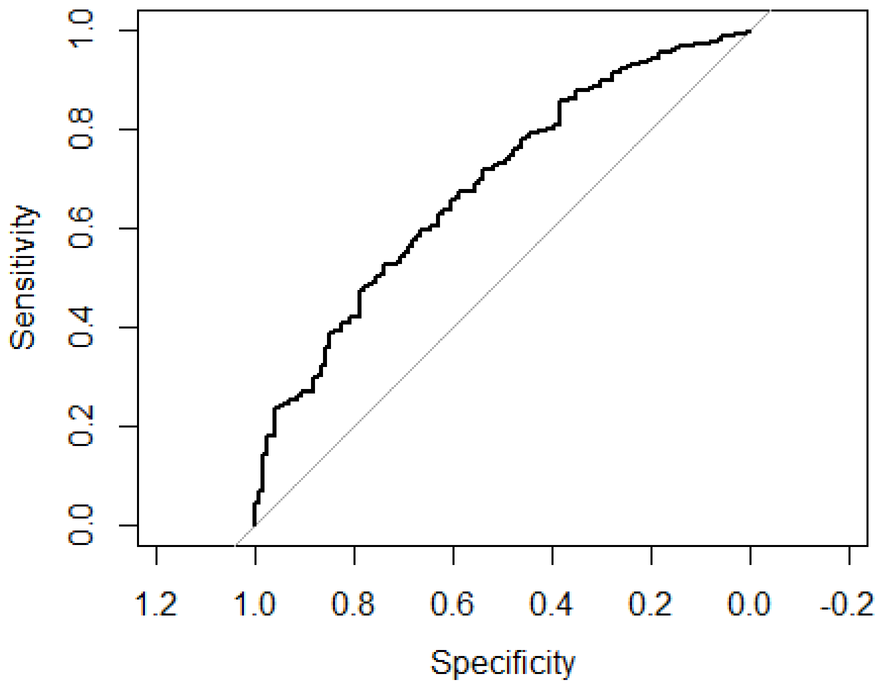


The area under the curve of the Credit Response Lift Chart is not very large, meaning that the model is not extremely useful in predicting good credit; however, it is still above the baseline, meaning that it does hold a small amount of significance.

The decile-wise lift chart for this model does not show significant gains in the earlier percentiles, indicating that the variables selected for the regression are not very significant to predicting good credit.

ROC Curve for Logistic Regression

```
#Performing a ROC curve and finding the area under the curve will continue to
give a better idea of the discrimination ability of the model.

gcreditroc <- roc(gcredit.valid.log$RESPONSE, gcredit.lreg.pred)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

#Area under the Curve:
auc(gcreditroc)

## Area under the curve: 0.6835

plot.roc(gcreditroc)
```

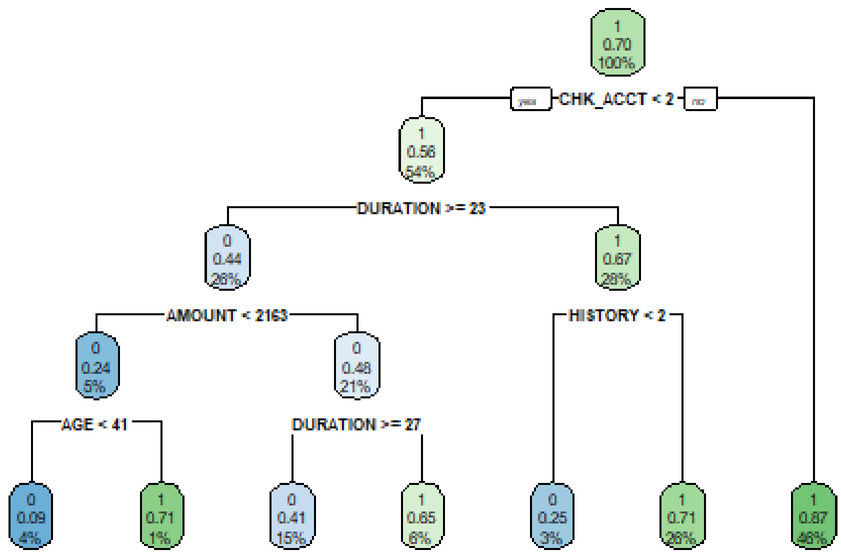


The ROC chart has an area under the curve which is slightly higher than the baseline of 0.5. This indicates that the model is, on a small level, good at predicting whether or not a user will have good credit.

Classification Trees (Default)

```
#The second part of the analysis is creating a classification tree to help de
termine a clear visualization of what factors cause a RESPONSE score of 1.  T
he default classification rpart() is used.

library(rpart)
library(rpart.plot)

set.seed(1)

gcredit.ct <- rpart(RESPONSE~., data=gcredit.train, method="class", control=r
part.control(maxdepth=4))

gcreditplot <- rpart.plot(gcredit.ct, main="Classification Tree")
```

Classification Tree




```
#Using the rpart.rules function to create rules from the given classification
tree.
gcreditrules <- rpart.rules(gcredit.ct)

gcreditrules

##  RESPONSE
##      0.09 when CHK_ACCT <  2 & DURATION >=       23 & AMOUNT <  2163
& AGE <  41
##      0.25 when CHK_ACCT <  2 & DURATION <  23                        & HIS
TORY <  2
##      0.41 when CHK_ACCT <  2 & DURATION >=       27 & AMOUNT >= 2163
##      0.65 when CHK_ACCT <  2 & DURATION is 23 to 27 & AMOUNT >= 2163
##      0.71 when CHK_ACCT <  2 & DURATION <  23                        & HIS
TORY >= 2
##      0.71 when CHK_ACCT <  2 & DURATION >=       23 & AMOUNT <  2163
& AGE >= 41
##      0.87 when CHK_ACCT >= 2
```

Analysis

To analyze the relationship of various predictors to whether or not someone has a good credit rating, two predictive analyses were conducted on the GermanCredit.csv dataset. After initial data exploration, the data was partitioned with 60% of the data in a training set and 40% in a validation set.

Variables were selected subjectively, with the goal in mind being the selection of predictors that would most affect the RESPONSE variable. The variables chosen for the logistic regression were DURATION, AGE, AMOUNT, and NUM_DEPENDENT, with relationship to RESPONSE. These variables were chosen because they are non-categorical.

Logistic Regression

The logistic regression performed gave the following estimated logistic equation:

$$Logit(Response = 1) = -0.4756 - 0.03876Duration - 0.09574History1 + 1.103History2 + 1.278History3 + 1.888History4 - 1.049Education1 - 3.717e - 05Amount + 0.01905Age + 0.2173Num\_Dependents.$$

Figure 1

Another look at the results of the logistic regression.

```
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.756e-01  6.436e-01  -0.739 0.459904
## DURATION       -3.876e-02  9.992e-03  -3.879 0.000105 ***
## HISTORY1        9.574e-02  5.832e-01   0.164 0.869605
## HISTORY2        1.103e+00  4.611e-01   2.391 0.016801 *
## HISTORY3        1.278e+00  5.346e-01   2.390 0.016854 *
## HISTORY4        1.888e+00  4.897e-01   3.856 0.000115 ***
## EDUCATION1     -1.049e+00  4.011e-01  -2.615 0.008911 **
## AMOUNT          3.717e-05  4.426e-05   0.840 0.401041
## AGE             1.905e-02  9.489e-03   2.008 0.044683 *
## NUM_DEPENDENTS  2.173e-01  2.696e-01   0.806 0.420250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the summary of the logistic regression, the variables with the highest significance were DURATION*** and HISTORY4***. EDUCATION1** has a slightly lower p-value and statistical significance.

The negative coefficient of DURATION*** suggests that the greater the duration of the account, the less likely the credit score is good. Due to the interpretation of categorical variables, the coefficient of HISTORY4*** is found by subtracting the given value from the intercept (Unknown, 2013); therefore, the coefficient is 0.2945 – 1.888 = -1.593. This negative coefficient indicates that the less likely it is that the account is critical (HISTORY4 = 1), the more likely the credit score is good.

Lift Chart & ROC Curve

The lift chart and ROC curve indicated that the model using the variables selected were not very significant for predicting whether or not a customer would have good credit. However, the lift chart did rise above the baseline given and the ROC curve's area under the curve was still greater than 0.5, meaning that the model was still of a little bit of use, if not much.

Classification Trees

When the classification tree method was chosen, using all variables instead of carefully selected ones for the logistic regression, the default model showed other variables that were also significant to predicting good credit.   This model indicates that if CHK_ACCT > 2, or if the money in the checking account is higher than 200 DM (Shmueli, 2018, p. 507), then the likelihood of them having good credit is 46%. However, if your CHK_ACCT < 2, meaning that the sum of money in the account is less than 200 DM, if DURATION <= 23 years, and if HISTORY > 2, then there is a 26% chance that your credit score is good.

Figure 2

Looking more closely at the rules created by rpart.rules.

```
gcreditrules

##  RESPONSE
##      0.09 when CHK_ACCT <  2 & DURATION >=       23 & AMOUNT <  2163
& AGE <  41
##      0.25 when CHK_ACCT <  2 & DURATION <  23                        &
HISTORY <  2
##      0.41 when CHK_ACCT <  2 & DURATION >=       27 & AMOUNT >= 2163
##      0.65 when CHK_ACCT <  2 & DURATION is 23 to 27 & AMOUNT >= 2163
##      0.71 when CHK_ACCT <  2 & DURATION <  23                        &
HISTORY >= 2
##      0.71 when CHK_ACCT <  2 & DURATION >=       23 & AMOUNT <  2163
& AGE >= 41
##      0.87 when CHK_ACCT >= 2
```

Using the rpart.rules function, other rules can be determined: there is a 75% chance your credit score is good when CHK_ACCT <2 & DURATION >= 23 & AMT < 2163 & AGE >= 41. Other, similar rules outlined in the same manner are predicted by the rpart.rules function.

Overall, the variables that were deemed most important by the default classification tree method were CHK_ACCT, DURATION, AMOUNT, AGE, and HISTORY.

Conclusion

The purpose of this analysis was to be able to understand the effect of various factors on whether or not a person's credit was good at a particular German bank. The methods used, logistic regression and classification trees, enabled a deeper understanding of both methods, especially the method of analyzing categorical variables, and partitioning data to produce validation results.

With this understanding, it will be possible to undertake more complicated projects and more accurately create predictions from datasets. This is an invaluable skill for data analysis and for creating business intelligence for successful companies.

References

Kassambara, A. (2018). *Logistic regressions essential in R: Classification methods essential*. STHDA. http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

Milborrow, S. (2020). Plotting rpart trees with the rpart.plot package. http://www.milbo.org/rpart-plot/prp.pdf

Unknown. (2020). *An introduction to logistic regression for categorical data analysis: From deviation to interpretation of logistic regression.* Towards Data Science. https://towardsdatascience.com/an-introduction-to-logistic-regression-for-categorical-data-analysis-7cabc551546c

Sheehy, R. (2018). *Understanding Confusion Matrix in R.* DataCamp. https://www.datacamp.com/community/tutorials/confusion-matrix-calculation-r

Shmueli, G., Bruce, P., Yahav, I., Patel, N., & Lichtendahl, K. (2018). *Data mining for business analytics: Concepts, techniques, and applications in R.* John Wiley & Sons, Inc. https://csuglobal.redshelf.com/book/read/1027801/?course_id=182329

Unknown. (2013). *R: plotting decision tree labels leaves text cut off.* Stack Overflow. https://stackoverflow.com/questions/16426007/r-plotting-decision-tree-labels-leaves-text-cut-off

Unknown. (2013). *Significance of categorical predictor in logistic regression.* Stack Overflow. https://stats.stackexchange.com/questions/60817/significance-of-categorical-predictor-in-logistic-regression